UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

Before The Honorable Susan van Keulen, Magistrate Judge

IN RE GOOGLE GENERATIVE AI COPYRIGHT LITIGATION) No. C 23-03440-EKL

San Jose, California
Wednesday, June 18, 2025

TRANSCRIPT OF PROCEEDINGS OF THE OFFICIAL ELECTRONIC SOUND RECORDING 10:05 - 11:26/11:43 - 11:49 = 1 HOUR 27 MINUTES

APPEARANCES:

For Plaintiffs

Joseph Saveri Law Firm, LLP
601 California Street
Suite 1000
San Francisco, California 94108
BY: CHRISTOPHER K.L. YOUNG, ESQ.

Bleichmar Fonti & Auld, LLP
75 Virginia Road
2nd Floor
White Plains, New York 10603
BY: GREGORY MULLENS, ESQ.

Blechmar Fonti & Auld, LLP
1330 Broadway
Suite 630
Oakland, California 94612
BY: ANNE K. DAVIS, ESQ.
LESLEY E. WEAVER, ESQ.

(APPEARANCES CONTINUED ON NEXT PAGE)

For Defendants:

Wilson Sonsini Goodrich & Rosati
95 S state Street
Suite 1000
Salt Lake City, Utah 84111
BY: PAUL SAMPSON, ESQ.

Wilson Sonsini Goodrich & Rosati
650 Page Mill Road
Palo Alto, California 94304
BY: MAURA L. REES, ESQ.
QIFAN HUANG, ESQ.

Wilson Sonsini Goodrich & Rosati
701 Fifth Avenue
Suite 5100
Seattle, Washington 98104
BY: ERIC P. TUTTLE, ESQ.

Transcribed by: Echo Reporting, Inc.
Contracted Court Reporter/ Transcriber
echoreporting@yahoo.com

Wednesday, June 18, 2025 10:05 a.m.

P-R-O-C-E-E-D-I-N-G-S

--oOo--

THE CLERK: Now calling case number 23CV03440EKL, case In re Google -- I'm sorry, In regarding Google Generative AI Copyright Litigation.

If Counsel can state their appearances, starting with the Plaintiffs.

MS. WEAVER: Good morning, your Honor. Lesley Weaver, Bleichmar Fonti, for the Plaintiffs.

THE COURT: Thank you, Ms. Weaver. Good morning.

MS. WEAVER: Good morning.

MR. MULLENS: Good morning, your Honor. Greg Mullens, from Bleichmar Fonti, here for the Plaintiffs.

THE COURT: Mr. Mullens, hello.

MS. DAVIS: Good morning, your Honor. Anne Davis, also from Bleichmar Fonti, for Plaintiffs.

THE COURT: Welcome. Ms Davis.

MR. YOUNG: Good morning, your Honor. Christopher Young, Joseph Saveri Law Firm, for the Plaintiffs.

THE COURT: Thank you, Mr. Young. Welcome.

MR. TUTTLE: Good morning, your Honor. Eric Tuttle of Wilson Sonsini, here for the Defendants.

THE COURT: Mr. Tuttle, hello.

MS. REES: Good morning. Maura Rees from Wilson

Sonsini for Defendants.

THE COURT: Ms. Rees, good morning.

MR. SAMPSON: Good morning, your Honor. Paul Sampson from Wilson Sonsini for the Defendants.

THE COURT: Thank you, Mr. Sampson.

MR. HUANG: Good morning, your Honor. Qifan Huang from Wilson Sonsini, also for Defendants.

THE COURT: Excellent. Mr. Huang, good morning.

All right. Thank you all for that. I think I've got everybody checked in. As I mentioned to my CRD, I'm sure you were thrilled to get to use your business cards today. I don't think anybody ever gets to use those anymore. Let me just -- give me just a moment. I just want my law clerks to know I don't have Teams, unfortunately. Whatever is on here is an old version, so I'm just going to open Outlook, if they need me to -- so we'll do it the old fashioned way. If you need to communicate with me to my clerks, you can do so via Outlook. I'm -- probably will recess at some point, so --

All right. We have a fair amount of ground to cover, so we're going to start, although Ms. Weaver is anxious to bring me up to date on something.

MS. WEAVER: I did just want to -- good morning, your Honor. Lesley Weaver, on behalf of the Plaintiffs. I wanted to let you know that we took your -- the spirit of

your order to heart, and we got a counteroffer on the custodian issue from the Defendants yesterday, and we all arrived here early to have further discussions.

THE COURT: Fantastic.

MS. WEAVER: And Mr. Mullens -- my firm will address that --

THE COURT: Okay.

MS. WEAVER: -- for our side, and Mr. Young will address source code, depending on what your Honor wants to hear and in what order.

THE COURT: Okay.

MS. WEAVER: Okay.

THE COURT: All right. Well, I don't want to bury the lead, so why don't we go ahead and start with custodians, and let's just see where we are on that. I was going to start with source code, but let's start with custodian.

Mr. Mullens?

MR. MULLENS: Good morning again, and Greg Mullens for Plaintiffs. Judge, as my colleague just mentioned, this morning, we met and conferred with the Defendants. And let me just get right to it and give you some numbers --

THE COURT: Okay.

MR. MULLENS: -- so you have them for custodians. Google's, sort of, ask initially was 13 custodians, Judge,

and yesterday they countered with also adding seven to that number --

THE COURT: Okay.

MR. MULLENS: -- so that would be 20.

THE COURT: Okay.

MR. MULLENS: You know, it took us time to consider that. We had a fair ask, a large ask in terms of our custodians. We certainly submit it was warranted, considering the size and the scope of the case, but we wanted to substantially compromise, as the Court ordered. Upon reflection and consideration and speaking with Defense counsel this morning, we have, from our ask of custodians, cut 20 of our custodians. We now stand at the 13 Google custodians, plus 20 more, which would be 33.

And in light of the two declarations that were submitted in support of the source code briefing, Judge, and those are two individuals who are never identified as a Google custodian, clearly very relevant to this case, we're also asking that those two individuals be added. So our number, just to give you a number, is 35.

I will say it took a lot for us to reduce our list down by 20 custodians. We're looking to -- or we are looking to stay there as our number. We want to engage with Google if they want to on it. We've done a lot of work in terms of our investigation into these custodians, also using Google's

documents which they produced to substantiate our ask. Your Honor accepted and considered for this dispute the organizational charts, which were actually very helpful in identifying unique custodians. And that's all to say that's where we stand at this very moment in terms of the numbers.

THE COURT: Okay.

MR. MULLENS: And we're looking to certainly answer any questions from the Court about the basis for our custodians or continue speaking with Google if that's what the Court wants.

THE COURT: Okay. Mr. Mullens, the seven that -- custodians that Google added in their most recent offer, were they seven off of your list?

MR. MULLENS: They -- it was unclear, and the answer is no, actually. They said they would offer up seven. They didn't give the names of who. They said they wanted to choose four, and we would choose three. We pressed them on who were at least the four that you were selecting -- from our list. It would be from our list of custodians.

THE COURT: I see.

MR. MULLENS: But they wouldn't give us the names. And at this juncture, part of our ask for our custodial ask is that we be the ones who pick the custodians and not sort of have a blind decision being made by Google as to who is

relevant based upon what we've done here in our compromises. So, right now, it's seven from our list of custodians from Google, but no indication who they are.

THE COURT: Okay. And where are the parties -- have the parties discussed search terms, or is there a plan once custodians are identified? Remind me where you are in the process.

MR. MULLENS: Yeah, search term negotiations are over for both sides.

THE COURT: Okay.

MR. MULLENS: So we're done with that.

THE COURT: Okay. So you have the terms in place.

MR. MULLENS: We do, Judge.

THE COURT: You're just trying to figure out --

MR. MULLENS: Yeah.

THE COURT: -- the pool. Well, that's a relief. Okay.

And I did look at the chart. Google DeepMind. I'm sure this was explained to me somewhere, but I had a lot of material to review. What does that refer to?

And don't worry, Google. I will hear from you too. I know why I'm asking the Plaintiffs.

MR. MULLENS: It really is the organization within Google that is responsible for generative AI --

THE COURT: Okay.

MR. MULLENS: -- development training (indiscernible), but that's --

THE COURT: So that's the umbrella, if you will.

MR. MULLENS: Yeah, that's the -- sort of the core entity or unit that handles it.

THE COURT: Okay. And then the products, some of which are still disputed in the motion to dismiss, those are all under or in the DeepMind --

MR. MULLENS: We would say they are, yes, Judge.

THE COURT: -- project? Okay. All right. That's helpful. Thank you.

From Google, anything to add to the update from Mr. Mullens?

MR. SAMPSON: Yes.

Good morning, your Honor.

THE COURT: Good morning.

MR. SAMPSON: Paul Sampson on behalf of Defendants. And just a couple of things that I wanted to kind of run through in response to Mr. Mullens' presentation.

In their brief on page five, they cite the Tremblay v. OpenAI case as an exemplar of the number of custodians. In that case, it was 24. We have come forward with the number 20 that we think is a good number, a ceiling on sort of what is needed for purposes of where we are in the case. And

what I mean by that is, you know, as I think the Court is aware, Judge Lee ordered the parties to focus on class certification issues. And it's been a real struggle to get there. We've been -- you know, I'm sure you saw in a number of the filings that we've had a lot of meet and confers to really try and focus the parties on, "Okay. You've asked for the universe. Let's focus this in on what's really needed for class certification." And the custodians issue is no different, you know, they came in asking for 62, you know, the triple of what's seen in other cases. And those cases are situated differently than us. They don't have class certification coming first. And so that was a big discussion and a lot of meet and confers to really narrow us down.

We have identified -- as Mr. Mullens said, we came forward with a compromise. We heard you that we should substantially compromise and to come forward with 20 custodians total. So that's seven additional from their list, and we offered, you know, "We'll pick four, you pick three." You know, we're willing to figure out what's the best way to identify that. I don't think it makes sense for them to pick all of them.

THE COURT: Why not?

MR. SAMPSON: Well, because the -- or the -- the case law talks about the responding party has, you know, the

ability to investigate and, you know, determine who the right witness is, and so we've come forward and put forward custodians based on that.

In response to the question about Google DeepMind, you know, originally, there were a couple different units within Google, Google Brain, Google DeepMind, that merged, and there's been a lot of reorganization since, and so not all of the products fall within Google DeepMind --

THE COURT: Okay.

MR. SAMPSON: -- in response to that question.

THE COURT: Do the four products that were part of the TRO, were those in DeepMind? It's not a -- it's not a pop quiz.

MR. SAMPSON: Yeah, yeah, I don't know the answer, so I'm going to defer to my colleague.

THE COURT: All right.

MR. TUTTLE: Your Honor, this is Eric Tuttle from Wilson Sonsini.

THE COURT: Thank you.

MR. TUTTLE: I want to be careful because I don't want to get anything wrong, but I think products -- like, products that are made available to users and public can tend to exist in different organizational units. Google DeepMind, in my understanding, tends to contain the sort of core model development functionalities.

THE COURT: Okay.

MR. TUTTLE: So there are models that are developed, and then those models are incorporated into --

THE COURT: Into products.

MR. TUTTLE: -- into products and services. Those products and services, I'm sorry, exist in other departments, organizations, and teams.

THE COURT: Okay.

MR. TUTTLE: Does that make sense?

THE COURT: Yes. Yes, it does.

MR. TUTTLE: Okay. Sorry.

THE COURT: So I think that answers my question. Thank you.

Mr. Sampson.

MR. SAMPSON: Yeah.

THE COURT: Anything further?

MR. SAMPSON: Yeah, another point -- you had questions about search terms and what's being done already. For the 13 custodians, we've already produced a number of documents, applying the search terms. We're, you know, cranking through those. We have 50 document reviewers that we've trained and who are ploughing through those documents at the moment, and the goal is to get documents produced for those individuals before the August 8th class certification motion.

THE COURT: Well, it's going to have to be much faster than that, but I understand, and I appreciate the parties' efforts, the discussions, and your efforts even up to and including this morning.

Okay. I had reviewed with some care the custodian issues, including the updated joint chart that had both sides' descriptions. I did take judicial notice of and look at the org charts, and I'm -- again, I'm pleased and appreciative of the effort that the parties have made. Where I came out was, again, looking at the products that are likely in the case, and I know you're waiting on the motion to dismiss order, and if there's a -- you know, a -- sort of say, "Okay. Well, let's call it six products." There are four that are -- were in from the TRO. There are others that are obviously being debated through the motion to dismiss, if a couple of those get through -- that's how I get to sort of six products. You need a couple of custodians as to each, again in addition to the work at least that had been done at the time I was looking at this.

So where I came out in terms of a number, that's my process behind it. It's not completely arbitrary. It will feel that way to each side, I am sure, but the number I came out with was 25 custodians, 25 from the list. If Plaintiffs now want the two declarants, those will count as part of the 25. And that gives you -- we've got the 13 now, so that's

12 to go, and I think it does make sense for Plaintiffs to have the laboring oar in selection because they have to make the most whittling down, if you will, from their initial ask. I will say, in reading those details, there was certainly a lot of overlap. And I appreciate that the descriptions are necessarily at a fairly high level, but there was a fair amount of overlap, and it certainly appeared to be. And as to certain products, it seemed there was almost a drill down of everyone who had touched it. So I am confident that there is room to further narrow and focus the search, particularly for purposes of class certification. So it will be 25 custodians. That's 12 additional, up to the Plaintiffs if they want the two declarants included in that. And beyond that, Plaintiffs can identify a remaining or an additional seven, and Google can identify an additional three. If the Defendant -- excuse me, if the Plaintiffs do not want the custodians, then they can use those two picks to further out for the additional -- for the additional 12. All right. So that gets us to a total of 25. Now let's talk about production.

MR. TUTTLE: May I ask one question, your Honor?

THE COURT: Yes, you may.

MR. TUTTLE: Eric --

THE COURT: You have to identify yourself for the record.

MR. TUTTLE: Sorry, I --

THE COURT: That's fine.

MR. TUTTLE: Eric Tuttle again from Wilson Sonsini.

THE COURT: Thank you, Mr. Tuttle.

MR. TUTTLE: I wanted to ask if it would be possible to, within the -- as I understand, we're adding 12, right?

THE COURT: Yes.

MR. TUTTLE: Within those 12, to have some sort of prioritization by about half for each because we want to make sure to produce the ones that are most critical for class certification before the August 8th deadline. And I'm -- I fear, right, that -- you know, we're well underway on the 13 custodians that have been agreed, but I fear that adding 12 may make it difficult, and I think not all of these custodians I would expect are essential for class certification. Some of them may be more important for the merits, and the discovery period will continue thereafter.

So I -- if we could get a -- if we could get an agreement that there was some rank order in terms of prioritization for class certification, that would help us ensure that we can focus discovery efforts on those that are actually needed for -- most urgently in the case.

THE COURT: Well, it is in Plaintiffs' interest to

prioritize their picks for class cert as well.

But let's talk about production --

MR. TUTTLE: Okay.

THE COURT: -- because it has to speed up substantially, in light of the class schedule or, yes, the class certification schedule that both sides have been aware of and have heard Judge Lee's views on. And we've been through this before with these parties -- excuse me, not with these Plaintiffs, but with Plaintiffs' Counsel with Google, and it is always a challenge, and I appreciate that, because there's a lot of material to review, but production is going to have to speed up, and that may mean you don't have time to review everything that comes out from the search terms, and the parties will have to formulate their stipulation with regards to rights to clawback and qualification under Rule 502(b), but -- and, obviously, you know, I mean that covers any privileged materials, anything that is inadvertently produced, and it will be inadvertent for purposes of 502(b), as well as meeting the other requirements. The reason for that is that, as a practical matter, it is -- there simply may not be time to review as the documents are produced. And I don't know what the custodians have and what the volume is, and I know that that makes Defendants uncomfortable, but it is the practical way that production has to proceed. And with the appropriate

stipulation and safeguards in place, then that should -- that shouldn't be a problem. If it's also as it relates to technical documents -- and, again, if you need to produce prior -- without review and you set that at your highest level of priority, subject to anything can be de-designated or lower designated later, then that is -- that's fine as well, because the Defendant needs these protections in place. But this production is going to have to get underway. I'm glad that it's already -- sounds like substantially underway as to the 13, so as to these additional, it's -- we've got to get it going. So rolling production will have to get started. We are already there, so -- we have to get started in two weeks and continue every two weeks. And the due date on the class cert motion is the 11th?

MS. WEAVER: I believe that's right, your Honor.

UNIDENTIFIED SPEAKER: I believe it's --

MS. WEAVER: I'm sorry, is --

UNIDENTIFIED SPEAKER: -- August 8th.

THE COURT: August 8th --

MS. WEAVER: August 8th, yes. Yeah, and we do have a 502(d) stipulation already in place, your Honor, and I would --

THE COURT: Okay.

UNIDENTIFIED SPEAKER: (Indiscernible).

THE COURT: This is Lesley Weaver for the Plaintiffs.

MS. WEAVER: I apologize.

THE COURT: Okay.

MS. WEAVER: Okay. Yes, we will work on a 502(d). I apologize for the misstatement.

THE COURT: That's fine.

MS. WEAVER: To date, we have received 5,000 documents --

THE COURT: Okay.

MS. WEAVER: -- and we produced 10,000 for the Plaintiffs, so we're going to move. And if your Honor -- I think, just to clarify, are you setting date production deadlines?

THE COURT: Yes, rolling production. We'll get started within two weeks and continue with the production every two weeks until the 29th of July, it'll need to be done, as to class cert. And you do owe the -- I'm speaking to the Plaintiffs. You do owe the Defendants a prioritization on these -- on the custodians.

MS. WEAVER: And we can do that. Is there a deadline by which you would like us give them the prioritization and --

THE COURT: Well, it certainly behooves you to do it as soon as possible.

MS. WEAVER: We'll do it this week. Okay. Thank you.

THE COURT: All right.

MR. TUTTLE: May I just be heard on two things related to that, your Honor.

THE COURT: Yes, Mr. Tuttle

MR. TUTTLE: So I understand the direction --

THE COURT: That's right. Mr. Tuttle.

MR. TUTTLE: Mr. Tuttle speaking, sorry.

THE COURT: That's okay.

MR. TUTTLE: I can't remember to do it.

THE COURT: Well, it's hard to have to keep saying it, but I appreciate the effort. Thank you.

MR. TUTTLE: I guess it means I'm getting up too often, your Honor, and I apologize for that.

But I just want to be heard on two things. One is we certainly hear the direction to get the -- we have been producing. We will make sure we continue rolling every two weeks and to get done by July 29th. I think -- I've got a lot of concern, I think naturally, with producing documents without a privilege review, and I think there would be issues with an order requiring us to do that -- I -- so I think we will look at the best way to accomplish the schedule that the Court has directed us to do. I do not understand the Court to be ordering us to produce without a

review. I understand the Court to be ordering us to figure out how to do this.

THE COURT: How to do it on this timeframe. And, often, what happens in this case is, as part of the review, you search for, obviously, counsel name in the documents or reference, but page by page review just won't be feasible on this schedule. And, again, the protections are in place, or the parties will get those in place, with regards to the 502(b) issues, so do your best.

MR. TUTTLE: We will figure out a way to get it done.

The other thing is, I just want to -- I don't want there to be a -- this needs to go both directions, I assume. And Ms. Weaver talked about the total number of documents that have been produced, but we've run through the Plaintiffs' -- you know, Steve Almond has produced 32 documents, which are 12 e-mail, so --

THE COURT: Those issues aren't in front of me today.

MR. TUTTLE: I understand. So we just -- but to be clear, this is -- everybody needs to get this done.

THE COURT: Everybody needs to get it done if it relates to class certification, which I imagine you have made some requests that go to that, directed to the Plaintiffs.

MR. TUTTLE: Yes.

THE COURT: And those documents need to be turned over.

MR. TUTTLE: Okay. Thank you, your Honor.

THE COURT: You're very welcome.

MS. WEAVER: And I apologize, but to clarify for the record, the protective order refers to the clawback procedure, and it was entered on April 4th. That's docket number 119 and paragraph 15 is -- deals with clawback and Rule 502(d). But we don't have a 502(d). We can endeavor to get that in place as soon as possible.

THE COURT: We have excellent lawyers on both sides. I'm confident you can meet and confer and get that done.

MS. WEAVER: Great.

THE COURT: Okay. That takes care of custodians. Now let's turn to the easy issue, shall we? So let me share with you my initial thoughts on source code. I don't have quite enough room up here, so let me get everybody organized. Okay. Source code.

First, let's address the rightness of the source code request. I reviewed the request for productions, the definitions that were used, the arguments on both sides as to what was included in which requests. The initial RFPs do not ask for source code where examples of documents that are

included in the requests are given. And there are many examples of documents and communications, but source code is never listed or addressed there. Obviously, the second set of RFPs does expressly address source code. It was designed for that. Google has made its objections to production of source code very clear. Our record is robust. And so for judicial efficiency, we're going to proceed and address it now. I'm not quite sure where the response date was, if we're still ahead of it.

UNIDENTIFIED SPEAKER: We're still ahead of it.

THE COURT: That's what I -- that's what I thought. But we are going to go ahead and address the issue now. And I'm taking Google stated objections with regards to this dispute in the joint statement and the declarations, et cetera, as -- that's the record, and I'm taking those as objections. They are noted, and we're going to work our way through those here today.

Where I always start in these cases is with the class definition, because, of course, when we're talking about source code, I am looking at relevance, and I'm looking at necessity. As I understand it from Judge Lee's order where she was striking the fail safe allegations, she said, you know, here's a definition that works, "All persons in the US who own a US copyright that was used by Google to train Google's generative AI models during the class period." If

there's something else that the parties are working from, I want to hear about that. And Mr. Tuttle is shaking his head no on behalf of Google.

And, Plaintiffs, is that our working class definition?

MR. YOUNG: I believe so, your Honor, with the exception --

THE COURT: Identify yourself, please.

MR. YOUNG: Oh, sorry. Christopher Young, Joseph Saveri Law Firm, for Plaintiffs.

That sounds about right, with the caveat that we've endeavored to limit the discovery that we're seeking, including the -- the source code we're seeking to inspect, to certain models and exemplar versions of those models.

THE COURT: I understand, but this is our working class definition.

MR. YOUNG: Correct.

THE COURT: Is that correct?

MR. YOUNG: Correct.

THE COURT: Okay. And what is the class period?

MR. YOUNG: We contend that it begins from 2017. The precise month and date, I don't recall offhand, but there is, I think, a putative or pending dispute about that temporal limitation. I understand that my friends at Google might contest whether or not that is an appropriate class period, but we -- as alleged, it's 2017 onwards.

THE COURT: Okay. All right. We'll start with as alleged. All right.

And the products that I understood, again from Judge Lee's tentative, where she was was to allow for Bard, Gemini, Imagen, and -- how do you pronounce LaMDA (pronouncing)?

MR. YOUNG: LaMDA (pronouncing).

THE COURT: LaMDA (pronouncing), thank you. Like the lambda.

MR. YOUNG: Like the Greek letter.

THE COURT: Like the lambda. Okay. Got it. Thank you.

And then others were somewhat on the fence. And I appreciate or I noted the Plaintiffs' request from those for all the various versions, which seems to expand the list substantially. I understand the thinking behind that.

So where I start is, it certainly appears that a relevant question for class certification is, what was Google sourcing -- harvesting from the internet? And we're obviously focused on data from the internet in the relevant time period. And how -- and then, ultimately, how was that used?

So my -- you know, of course, the overarching question is, is source code the best indicator of that? And I've read through the declarations taking issue with where URLs

are shown, et cetera, but my first question is, really, what has Google produced to date that shows the -- well, we'll start with shows the content of training sets. Now, I know we have the -- what was harvested from the internet, and then what actually ends up in training sets, and then how those are used. But what has Google produced to date that shows what actually -- what -- the content, if anything, that shows content of training datasets?

And, Mr. Wong (sic), respectfully, let me get Google's answer. And I'm going to hear argument from both sides, but I want to get some basic elements in before me.

MR. YOUNG: Yes. And, your Honor, just to nip this early before it -- it's Mr. Young. Thank you.

THE COURT: Well, Mr. Young, I apologize. I apologize.

MR. YOUNG: No problem.

THE COURT: I'm going to put my Plaintiffs where my Plaintiffs belong. Okay.

MR. TUTTLE: Thank you, your Honor.

THE COURT: Mr. Tuttle.

MR. TUTTLE: So the --

THE COURT: Mr.?

MR. TUTTLE: Sorry. I will get this. Mr. Tuttle speaking.

The -- as we've -- we've had a number of -- we've had

many discussions with the Plaintiffs about the training data, and basically it's an extraordinarily large amount of data. And so we proposed from the outset that what we do is we provide the Plaintiffs with information describing the datasets used for each model, and then Plaintiffs select representative or sample datasets. We eventually agreed at some level on doing so, and we got a -- we got a set from them. We provided these data cards that -- for models, for all the different models and all the different --

THE COURT: So what's a data card? What's on the data card? We're drilling down from datasets, I take it.

MR. TUTTLE: Well, I'm going to get to the datasets, but the --

THE COURT: Okay.

MR. TUTTLE: -- data card, you know, I'm using that as a general term. They vary, because we're talking about a long period of time and different teams, but at a high level, a data card describes a model and the training datasets that were used for that model.

THE COURT: Okay.

MR. TUTTLE: So it describes in English, right, with descriptive names. And the amount of detail and additional information varies, right, by model and time period and team.

THE COURT: So you said it identifies a model and

the training data used for that model?

MR. TUTTLE: Correct. It identifies the -- just -- to be clear, it identifies sets of data, so, you know, web -- like, it might identify a set as Wikipedia, or it might identify a set as data crawled from the internet, or it might -- you know, and so on. And then it will -- it will -- and then it will -- and it'll have additional information. So we provided this to the Plaintiffs, so that they could select datasets that would be of interest to them, because there are, we assume, some datasets that are of more interest to the Plaintiff than others. The Plaintiffs made a selection, and I don't have a date offhand --

UNIDENTIFIED SPEAKER: May 20. May 20.

MR. TUTTLE: May 20th, we were given the datasets that Plaintiffs selected.

THE COURT: Okay.

MR. TUTTLE: And we have -- we -- this is another -- whole other issue that I don't want to spend time on, but we -- it's too large, it's extremely large, and it's also very sensitive. So what Google offered to do is to make it available in a review environment within Google --

THE COURT: A clean room environment --

MR. TUTTLE: Well, it's a remotely accessible environment. It uses Google's cloud technology, right, to make an environment in which this data will be placed and

where the Plaintiffs and their experts, under the protective order, can securely connect remotely using a Chromebook. And they will be able to access this environment where all the data will be, and they will be able to operate on it and explore it, right? And --

THE COURT: And you're talking about just one of the selected datasets?

MR. TUTTLE: All of the selected datasets --

THE COURT: Okay.

MR. TUTTLE: -- will be -- that can -- you know, that -- all of the selected datasets that, in fact, correspond to a model -- like, we're working through it. I think in some cases, they may have picked a dataset that was not actually used with any of these models. But to the extent the dataset was correctly identified as a dataset used with the model, the dataset is being located, is being moved into this environment, so that the Plaintiffs can review it.

THE COURT: Uh-huh.

MR. TUTTLE: Even this limited number of datasets is very, very large, so we are in the process -- like, I think about half of it is currently in there, and the environment has been built. This was slowed down a bit by disputes between the parties about the nature of the environment, but we went ahead -- given the timing, we just

went ahead and built it and put the data in there. And I think we have -- we'll be shipping the Chromebooks to the Plaintiffs this week, and we just need to get the account information, and I think we should be able to get them access, so --

THE COURT: And they can access it remotely?

MR. TUTTLE: They can access it remotely. They can access it from anywhere. It will be -- I mean, it will be -- they need to be on one of these laptops.

THE COURT: I understand.

MR. TUTTLE: They need to be one of these credentialed users. And there are various security measures in place. So the data is being made available there. And I think the Court is quite right to focus on the training data, because the training data is going to tell us what works are -- were used to train the models.

THE COURT: So -- and will the training data, does it identify specific works or specific content?

MR. TUTTLE: The training data contains the content.

THE COURT: Okay.

MR. TUTTLE: And, typically, it contains metadata -- and, generally, it contains metadata about the content, which -- for example, for web crawl data that Google crawled, it should generally contain the URL. There are,

you know, thousands and thousands and thousands of datasets across many teams over many years, so we haven't been able to look at all of it and determine if that is always the case. And if it should turn out that there's information that is not present, we can figure out a way to deal with that. But that information, in any event, will not be in the source code.

THE COURT: But your understanding, your representation to the Court is that the datasets reflect the training model content --

MR. TUTTLE: -- the content.

THE COURT: -- the content and the metadata.

MR. TUTTLE: Generally.

THE COURT: Generally. So why generally? Or just tell me what your --

MR. TUTTLE: I say that only because -- like, because we're -- it clearly contains the content, right?

THE COURT: I got that. I got that.

MR. TUTTLE: Because that's the content that is exposed --

THE COURT: I'm trying to get to the URL. So tell me how --

MR. TUTTLE: Yeah. So the URLs for, again -- what -- my understanding, we just haven't been able to review all of the training data, so -- because it's many, many -- as I

said, it's many, many thousands of files all over the place, but my understanding is that for -- as for, like, public web stuff that Google crawled and collected and used for training data, that the URL metadata should generally exist in -- the URL metadata should generally exist in the training data files. To be clear, these are the files that are the -- that are -- the training data that is available to the model to use to train.

THE COURT: Uh-huh.

MR. TUTTLE: There is some sampling that occurs during the training and so forth. And so it's -- it -- we can't say -- like, there are -- in some cases, we -- like, the -- what was actually used got saved, and so we would have a record of that. But in other cases, all we have is the final set of the data that was available to be trained on, if that makes sense.

THE COURT: What's the difference between those two?

MR. TUTTLE: So it could be that the training datasets that the model draws from to train --

THE COURT: Yes.

MR. TUTTLE: -- there could be data in there that was not sampled, but the data that could have been chosen is in those files.

THE COURT: Okay. All right. And you say you are

shipping Chromebooks this week?

MR. TUTTLE: Yes, I think they can go out.

UNIDENTIFIED SPEAKER: We just need to know who to ship --

MR. TUTTLE: Yeah, we just need to get the information on --

THE COURT: All right.

MR. TUTTLE: -- who to ship them to.

THE COURT: So -- and is there any reason why review in this environment, or, excuse me, why Plaintiffs' access to this environment can't be up and running next week?

MR. TUTTLE: I'm not aware of a reason. Oh, well -- there is an ongoing dispute about cost, and I --

THE COURT: Okay.

MR. TUTTLE: -- it has been teed up for a motion, but I don't know that it is necessary to resolve because, in our view, it is premature.

THE COURT: Okay.

MR. TUTTLE: Google has agreed to set this up at its own expense.

THE COURT: Okay.

MR. TUTTLE: And -- but it has -- and it has said that -- you know this environment --

THE COURT: I'll look forward to the motion --

MR. TUTTLE: Yes.

THE COURT: -- if the parties are unable to work it out.

MR. TUTTLE: Yes.

THE COURT: But it won't stop the Plaintiffs from accessing as early as next week?

MR. TUTTLE: No, it will -- from our perspective, it does not. We have offered to make it available up to -- I believe it's $1 million of commercial rates of free use. We did reserve the right to pause it if they go well past $1 million, but unless they hit $1 million in commercial rates, there's nothing -- there won't be anything stopping them from using it, from our perspective.

THE COURT: Okay. Well, as I say, I don't expect access to be limited on any cost issue at this time, and, obviously, if the parties can't come to an agreement, you'll come back to me, and we'll deal with it.

MR. TUTTLE: Yes. It's premature, because I hope we will never get to $1 million.

THE COURT: I hear you. I hear you. I hear you. I hope that's right, but we're talking about a lot of data, so different issue.

Okay. So I asked you to come to the podium, Mr. Tuttle, on behalf of Google, because my question was, so what has been produced? And it sounds like that -- again,

I'm pleased the parties had discussions and went to the sampling approach and that Plaintiffs identified datasets, and Google has endeavored to make those available through this environment, and Plaintiffs' access and inspection of that should begin as early as next week.

MR. TUTTLE: Okay.

THE COURT: Okay. Don't go away. Let me just --

MR. TUTTLE: I'm right here, your Honor.

THE COURT: Okay. That's very helpful. Thank you, Mr. Tuttle.

MR. TUTTLE: Thank you, your Honor.

THE COURT: Mr. Young.

MR. YOUNG: Thank you, your Honor.

THE COURT: Thank you for making sure I'm calling up the right person. It's your motion. Let me start -- let me hear from you. I appreciate the explanation from Google. That seems to get us a significant way down along the lines of your request for production 65. So, again, with the necessity for source code standard firmly in mind, what's the issue?

MR. YOUNG: Thank you, your Honor. So before I talk about why data cards --

THE COURT: Pull a microphone close to you, Mr. Young, just so we're sure we're getting the recording.

MR. YOUNG: Thank you. Is this better?

THE COURT: That sounds good.

MR. YOUNG: Great. So before I talk about the source code, I do want to respond to a couple of points about the training data and the training data protocol, because I would be remiss if we just didn't put that in the record. I mean, from our position --

THE COURT: Just slow down a little bit.

MR. YOUNG: Yeah. From Plaintiffs' position, one of the reasons why this has being, like, teed up for a brief at all is we've taken issue with how the training data has -- is going to be produced to us, just given that we've engaged in months of -- months of negotiations, and we feel like Google has foisted unilaterally their initial position, essentially wiping away months of effort, including on this cost issue. But I understand that that issue is not ripe, but I just would be remiss if we just let those statements stand without some sort of response here on the record.

THE COURT: Okay.

MR. YOUNG: Now, with respect to the question about why data cards and model cards are not enough, what that really answers is the what, not the how and the why. With respect to categories of source code we think are, you know, I wouldn't say essential, but are the best evidence of some of the conduct at issue, that will need to be shown at a class wide -- on a common class wide basis, or the source

code, as your Honor identified, have the organization and collection of data, so crawlers scraping from the internet, the processing of that data, so processing from raw data into a form that is ingestible by the machines, by the models for the training process, and at the end step, what is needed to prevent what is called leakage or regurgitation of training data from the model. And, you know --

THE COURT: So let's focus on what source code -- what your argument is as to the need for source code for purposes of class certification.

MR. YOUNG: Right.

THE COURT: Let's put aside the issue, to the extent we can, of fair use, and, you know, what I think of that is what happens to data once it's obtained. But if you have access to, "Here's the training data" and you're able to review that, again keeping the class definition in mind -- you're able to review that, you're looking for text or images or whatever that is subject to copyright and is included in a training set, you're also able, at least generally, using Mr. Tuttle's word -- should be able to look at the metadata as to where that came from. So if you have a URL and you have the data that ended up in the training set whatever -- in whatever form, doesn't that inform your issues regarding class certification for numerosity, typicality, and et cetera?

MR. YOUNG: Perhaps. And I say that because the data cards and the model cards we have received are essentially snapshots. The source code would really be the best evidence of all that data. And I point your Honor to paragraph 11 of the Liekti (phonetic) declaration that was submitted by Google. I mean, for example, that paragraph --

THE COURT: Hold on. Hold on.

MR. YOUNG: Yes.

THE COURT: Let me get there, Mr. Young.

MR. YOUNG: Sorry, your Honor.

THE COURT: That's all right. I'm going to keep slowing you down, so --

MR. YOUNG: Yeah. So whenever you're ready, your Honor.

THE COURT: Wait, let me -- I read this with some care. Where did it go?

MR. YOUNG: So Mr. Liekti -- and I believe this is the second sentence of paragraph 11 of his declaration -- says,

"The source code for training generally includes mixture files, so scripts that identify the file names and locations of the various compilation datasets that are used in the training process, as well as proportions by which these

datasets are sampled from."

So why is that important for class certification, right? What Mr. Liekti is saying, if I'm understanding him correctly, is that their source code that compiles the various training datasets that Google maintains into what is the -- what is going to be the training corpora of that model, right? So whether or not a training dataset is used or is not used for a training dataset, this code will tell us what exactly is used. Further, it will tell us what proportion that is being used, but we anticipate that one of the things Google may say is that a dataset may not -- a dataset containing copyrighted information may or may not have been used more or less, right, which may have implications, not only for fair use, but also for class certification and creating -- perhaps create individualized issues for datasets. So that's why we would need information about that source code. Right.

Setting that aside --

THE COURT: Hang on. Hang on.

MR. YOUNG: Yeah.

THE COURT: Let me just take a look at the language. Okay.

MR. YOUNG: Okay. So, you know, your Honor, if you reflect back on the, kind of, three-part framework of the three categories of source code identified early on in

my argument that we would argue are -- we would most like or would be most explanatory for class certification purposes, this is really the organization or processing point, because this is the organization or collating of data before it becomes used as training data for the models, right?

Now, taking a step back to the web crawlers. Now why is that important? The web crawlers, including kind of the filtering -- and this is described by Mr. Liekti in paragraph four of his declaration, as well as by Mr. Ibrahim in his declaration, identifying specific paragraphs.

THE COURT: I'm sorry, Mr. Young, you need to slow down.

MR. YOUNG: I apologize.

THE COURT: That's fine.

MR. YOUNG: I will take a breath. This is a bit of a problem for me, so I will endeavor to correct my cadence.

Now, if you'll --

THE COURT: So we were talking about -- you wanted to move back, if you will, to crawler source code.

MR. YOUNG: That's right. And the reason I am framing it this way is it's sometimes helpful to think about the life cycle of data as it gets fed into the model, right? So the crawling is really the first step --

THE COURT: Right.

MR. YOUNG: -- so this is the original copy. So besides identifying class works, what is -- why this is also important for class certification is because it also bears on common issues of intent and willfulness on behalf of Google, because the training dataset will not tell us what processes Google may have done at that initial collection stage to filter, include, or exclude data, right? You could imagine, right? And, you know, of course we haven't seen the source code, but even from a common understanding of how Google works, Google Search, which Mr. Liekti identifies in paragraph four as being the provenance of training data, has -- blocks or restricts access to certain websites. If Google has chosen to restrict or not chosen to restrict, for example, pirate websites that it crawls, that is going to be common evidence used by the class to -- that would be common evidence used by the class that goes directly to -- only to questions of infringement.

THE COURT: But if you have the URL data in the metadata from a training set, then aren't you aware of what's coming from a -- you know a legitimate website or a pirated website?

MR. YOUNG: Well, we also -- the issue is also we do not know what the -- how Google maintains its training data. For -- from my experience, having litigated other generative AI cases, it can be a mixed bag, but, generally -

- but we do know that the source code extracts URLs and text from the -- from those sources that Google crawls. And I would point you to Mr. Zhao's declaration. I think this is paragraph three of his declaration, where he admits that Google's crawlers, one of the functionalities is to extract URLs and content that are destined to be -- for use as training data. So we know that the source code evidence is that, and the -- well --

THE COURT: That's hotly disputed as to whether or not you can tell URLs from the source code.

MR. YOUNG: Yeah. Well, I would --

THE COURT: So tell me where you think that the Zhao declaration points to that?

MR. YOUNG: Yes. I will point you to paragraph three --

THE COURT: Uh-huh.

MR. YOUNG: -- and the second sentence beginning "these crawlers." So --

THE COURT: Hang on. Hang on. Hang on. It's actually the third sentence, but that --

MR. YOUNG: I apologize. The little -- this little second sentence kind of evaded my eyes there.

THE COURT: All right. I see the language about extracting URLs.

MR. YOUNG: Correct. You know, there's an

inconsistency I would admit within the Zhao declaration wherein he says in paragraph five -- and your -- if your Honor would -- I mean, I would be remiss if I didn't point that out -- say that the source code will not identify specific URLs, but this seems in direct tension with paragraph three of that declaration, where he says that it would extract URLs. And even so, if the -- if -- what Google has asked us to do is to pick datasets essentially in the dark with imperfect information, whereas the source code would give us a direct snapshot into what is used to organize and collate training datasets. We are being asked to select representative datasets from data cards and model cards, which are snapshots of various models and versions which we have very limited information about, whether they're production models, benchmark models, intermediate models. But the source code was really the best source of evidence for us to suss out what relevant datasets we really want and really need, and --

THE COURT: How would you possibly have time for that, given the vast amounts of code that you're talking about?

MR. YOUNG: So it is -- so, you know, there's a bit of a tension here, right, because our experts do not -- like, we -- taking at face value, right, that Mr. Liekti is correct that there are millions of lines of code, my

understanding from our experts is that a lot of that code is not any code that we are interested in at all. I mean, really from even just these declarations -- and I think our experts have identified, just from the scant document production, the code -- specific code that we are actually interested in, which I don't think Mr. Liekti takes much -- no, he quibbles with it, but I think they're actually more in line than they are in disagreement. We have identified specific code that we are interested in. For example, to the extent that they're -- you know, you would believe --

THE COURT: With what specificity when you say you've specified?

MR. YOUNG: Yeah. So Mr. Ibrahim has gone through some of the documents that Google has provided, with the understanding that was before the recent, kind of, custodial -- first wave of custodial documents, and identified specific scripts and code that were identified in the documents. I mean, he has identified something he believes is a scrape code, which we understand from Mr. Liekti's declaration to be something that Google uses to convert crawled data into training-ready data, right? And I don't think Mr. Liekti has disagreed with what Mr. Ibrahim says -- said what this is at all, but, really, rather than being scraped from the internet, it was scraped from something that was already corrected -- collected from the internet,

excuse me. Mr. Ibrahim has also identified code that identifies leakage and regurgitation and how that would "require indexing training data," which is why we're also interested in that last stage, because that indicates Google's knowledge and knowledge of whether or not the training datasets that it collected contain copyrighted works, which would again be common evidence that each member of the class would rely on. But without actually knowing what's in the code and what's referenced in the code, even if it's not looking at referencing specific works in the training dataset, just knowing, for example, "This is the dataset that we are indexing, and that's the dataset we want to look at," right, setting aside, you know, there might be a dataset, for example, called, like, e-mails, which we might not be interested, right? But just having the references or the identification of the datasets that Google is using in its source code and referencing in its source code would be tremendously helpful at class certification.

Now, your Honor was giving me a bemused look, so I will pause to see if I can offer some clarity.

THE COURT: No, I'm following your argument. I get the -- I get that you want the source code. I understand. I hear the argument of, "Well, it's the best evidence of." I'm trying to get to the necessity point. And, again, I'm focused on the class certification issues

and the identification of where did Google go to get data to build training sets? And it sounds -- and, again, if the datasets in the -- that are available for inspection contain the metadata, it seems like that goes a long way to identifying where they went, what they collected, and then, you know, obviously, there's issues around what actually ends up in the training data.

MR. YOUNG: Yeah. Well, your Honor, I think just to directly respond to that point, right, I mean, we have -- we have been -- you know, as -- I believe I heard Mr. Tuttle say, right, that even one of the representative datasets that we've identified from the data cards and model cards may not have been used by Google at training, which really highlights why the process we've been doing here now is an imperfect one, because we are being asked again just to kind of pick and choose, from the limited information we have, what the most relevant training datasets would be for class certification. And part of the difficulty we've been having is that we do not know for a fact which -- or know with any reasonable amount of certainty which datasets have been used for training the models or if any of the datasets themselves contain copyrighted works at all, because they are -- generally, they're not descriptive. We do not know like -- besides, like, the name --

THE COURT: No, but you're going to have a chance

to investigate --

MR. YOUNG: Of course. Of course.

THE COURT: -- them and look behind the curtain, if you will. Have the Plaintiffs asked or endeavored to ask Google as to when you got to the -- went through the sampling and selection of datasets, as -- was there a discussion with Google for confirmation as to whether or not the datasets identified by Plaintiffs were actually used in training?

MR. YOUNG: There's been a long back and forth about the training datasets. Even getting to the point of getting data cards in the first instance was a suggestion from Plaintiffs. So, your Honor, we've been working really, really hard to attempt to identify the training datasets. And, you know, this is the first for -- at least from my recollection, and perhaps my colleagues have heard this before, that even one of our own training datasets was not used. We've been simply asked to identify training datasets, and we provided them a list.

THE COURT: I'm sorry?

MR. YOUNG: We've been asked to provide a representative set of -- representative datasets --

THE COURT: Right. And have you asked Google to confirm that those datasets were actually used in training?

MR. YOUNG: Not -- no, I do not think we have --

THE COURT: Okay.

MR. YOUNG: -- but we -- based on -- I mean, based on the model dataset -- the data cards that we've had, these are the most -- at least, based on the names of the datasets, these seem to be the most likely ones to be -- contain class works.

THE COURT: All right. That's very helpful, Mr. Young. Thank you.

MR. YOUNG: Uh-huh. Now, your Honor, I do want to -- we've talked about crawlers and the, kind of, organization filtering scripts. I do want to touch briefly on the last category of scripts that I was referencing, which is --

THE COURT: Okay. Hang on. Hang on. Hang on.

MR. YOUNG: Yes.

THE COURT: Give me just a second. Okay. Go ahead.

MR. YOUNG: Yeah. So the last category of source code that we are seeking are these "mitigation-type filters."

THE COURT: Uh-huh.

MR. YOUNG: So these are scripts at the back end that check, essentially, for lack of a better word, what is coming out of the model. So, for example, this could be checking for PII, for example, if someone's, like, e-mail

address or someone's, like, address is about to be spit out by the model, it'll check. Why we think this is relevant for class certification is that this is just another source of relevant training data, because, for example, if -- because how we understand these things to work, based on the limited discovery we have, as well as just our experts' understanding -- broad understanding of how these generative AI models function, you cannot have one of these filters on the back end unless it references what you are trying to stop from coming out. So you can imagine --

THE COURT: Well, when you say on the back end --

MR. YOUNG: Yes.

THE COURT: -- where is it? Is it on the training data? Is it as the data moves from the crawler into the training dataset? Where?

MR. YOUNG: So this is on the very -- right before the model outputs information --

THE COURT: Uh-huh.

MR. YOUNG: -- for example, you could imagine the model just taking out of the copyright -- or maybe keep it in, right? It is about to emit, for example, two, like, sentences for Harry Potter or a chapter of Harry Potter. The filter will go, "Hey, wait, this is a copyrighted book. Don't do that. Instead, say this," right?

THE COURT: Uh-huh.

MR. YOUNG: Say, "I cannot emit copyrighted works."

THE COURT: Okay.

MR. YOUNG: Now, how we understand this to generally work -- and, you know, perhaps Google has devised a better way to do it, but based on the document that Mr. Ibrahim has identified that says that Google needs to index training data to develop this, we do not think that they do. That means that they have to develop the filter using copies of the copyrighted work that they are trying to stop from coming out --

THE COURT: Okay.

MR. YOUNG: -- which, of course, is common evidence of knowledge of -- that their original training corpus contains copyrighted work.

THE COURT: So how does that relate to the need for source code?

MR. YOUNG: The -- because this is evidence in the source code. It'll refer back to the -- either the training dataset or some training dataset of a collection of copyrighted works. And just the mere existence of this is also evidence of that intent. But the source code would be the best evidence of the specific concerns and intent that Google had about stopping copyrighted works from being

emitted.

THE COURT: Let me make a note here. Let me look at something. Thank you, Mr. Young. Let me see if I have another question for --

MR. YOUNG: Absolutely, your Honor. Thank you.

THE COURT: Okay. All right. That's very helpful. Thank you.

MR. YOUNG: Thank you very much, your Honor.

THE COURT: All right. From Google, I have some follow up questions. Who's got the mic on this one?

MR. TUTTLE: Thank you, your Honor. Mr. Tuttle.

THE COURT: Mr. Tuttle for Google.

So let's go to a couple of my questions. I will give you an opportunity to kind of backtrack and address the explanations provided by Mr. Young. So datasets -- sample datasets identified by the Plaintiffs. You've been working with those to get them in this environment. Has there been confirmation by Google, or is Google able to confirm whether or not those datasets were actually used in training?

MR. TUTTLE: Yes, we are able to do that, your Honor, and we do intend to -- we're working through the list, and we're intending to, you know, explain to the Plaintiffs what we've found. I think where the Plaintiffs used a data card for a specific model and chose a dataset described on that data card, that dataset was used, and we

have it. I think there were a few instances where Plaintiffs chose to use a different type of document, a document describing a dataset, and asked for that dataset as used in a model, and it was not. I -- that's my memory is there's one or two examples of that.

THE COURT: Uh-huh.

MR. TUTTLE: But if the Plaintiffs chose a dataset described on --

THE COURT: A data card.

MR. TUTTLE: -- a data card for a model, then that dataset is for that model, and we will -- and we will have found it, and we will have provided it.

THE COURT: Okay.

MR. TUTTLE: And if there -- if anything comes up, if there's any issue, we will disclose that to Plaintiffs, and we will work through it. But I -- the only ones I'm -- the only ones I can think of right now where I think we have not found that dataset in connection to that -- with that model was where the Plaintiffs pointed to a document that was not describing a particular model, right? It was describing a dataset or a project or something, and they said they wanted that dataset, and it -- that was not used in -- this is my memory, it was not used in connection with that model.

THE COURT: Okay.

MR. TUTTLE: And, again, I think the Court is quite right to focus on the class definition that the Court has -- that Judge Lee specifically has given the Plaintiffs' leave to amend to and once the motion to dismiss is ruled on, which is US persons who own a US copyright in any work used by Google to train Google's generative AI models. So what we're focused on are the works that were used to train, and the training datasets for the models will tell us, by what's in them, what those works are. And the source code does not tell us that.

THE COURT: Okay. Mr. Young made the point about the mitigation filters.

MR. TUTTLE: Yes, your Honor.

THE COURT: And that the source code has the instructions as to when and how to use those. So what -- does the training -- how do the -- it doesn't sound like the datasets or the data cards would contain that information.

MR. TUTTLE: Well, the purpose of the mitigation -- that mitigation software is to use the training data that was used for the models in order to create -- in order to be able to detect, you know, repetition or duplication of the training data for that model in the output of that model. So the code for these mitigation measures will contain the detailed algorithms for how an index would be built, how it would prevent recitation in the output of a model, but it

won't tell you -- it won't tell you the data, and the data -- and, again, the data is the training data for the model. That's the whole point, right, is to have the -- is to have those -- that mitigation software able to detect the presence of the same data that went into the model, right? Because it's trying to -- it's trying to block, you know, repetition of that data in the output.

THE COURT: So if I'm understanding your explanation, it's you have the content of the training data.

MR. TUTTLE: Yes.

THE COURT: That's what we've been talking about. You have the exact content. To the extent that there's a mitigation filter to limit the output from that training data, that's okay. But the training data is what it is.

MR. TUTTLE: That's correct, your Honor. And I -- you know, I just -- the source code isn't going -- again, all -- the source code will get into the gory, technical details of how things are done, but it will not, at the end of the day, tell us what works, are involved, or anything like that.

THE COURT: Uh-huh. Uh-huh. Uh-huh. Okay. I started you off by answering some of my questions. Did you have a further response to Mr. Young's presentation?

MR. TUTTLE: Well, I don't want to -- I would prefer to focus on the things, if anything, that concerned

the Court or had questions. I don't want to -- I don't want to spend time on issues that aren't important. I will just say, like, walking through the history of the crawling -- again, the source code for the crawling won't -- as the declaration said, it won't tell us what URLs were crawled. I don't think the Plaintiffs are asking for all the URLs that Google has ever crawled in the history of the internet. That's what Google does. Google tries to crawl as much of the internet as possible. That's how search works. But in any event, it doesn't relate to the class definition. The class definition isn't a class of all people whose copyrighted works were crawled by Google. If they -- it's all people whose works were used to train Google's generative AI models. So there's all these steps, right?

THE COURT: Uh-huh.

MR. TUTTLE: But at the end of the day, the training data that is used to train the model is the source for the works that are used for that purpose.

THE COURT: Mr. Young pointed to what he -- may be an inconsistency in the Zhao declaration, and that's docket 152-2, specifically with regards to the language in paragraph three. And I'm quoting from the third sentence of paragraph three,

"These crawlers handle everything from domain resolution and protocol

negotiation to parsing the robots.text protocol, extracting the URLs and content from web pages, and crawl scheduling and rate limiting."

And as I understood Mr. Young's argument, he was focused on the extracting URLs language and acknowledging in paragraph five that the declarant says that specific URLs of websites is not demonstrated in the source code.

So can you explain -- can those positions be harmonized or can you explain what -- is it Mr. Zhao or Doctor Zhao is referring to in paragraph three with regards to the extraction of URL?

MR. TUTTLE: Yes, I think that Doctor Zhao is explaining that the source code for the crawlers will contain the routines that do the extraction of the URL, but they don't -- but the URLs are not themselves in the code. So the -- when a Google crawler crawls the web, right, it extracts the content of websites and metadata about the websites, and then that information is stored, and -- but the code --

THE COURT: Is the instruction to do it.

MR. TUTTLE: -- is the instruction on how to do that extraction. It doesn't contain the extracted text or metadata, which -- such as the URL.

And then what matters at the end of the day is what

training data got selected and used to train the models, right? And so that's, again, I think why I keep focusing and the Court seems focused on that training data --

THE COURT: Uh-huh.

MR. TUTTLE: -- which is sort of the end of the line of what is -- of what the models are trained on.

THE COURT: Provided there's metadata that shows where it came from.

MR. TUTTLE: I understand, your Honor, and I believe there will be. If it turns out that there's some example where that's not present, then we will look to see where -- if there are -- if there's other information that's available that would contain it, but it would not be in the source code regardless.

THE COURT: Uh-huh.

MR. TUTTLE: So the source code is not -- I believe it will be in the training data --

THE COURT: I understand.

MR. TUTTLE: -- but if it turns out not to be, we'll have to figure out some other -- if there's something else. But in no event --

THE COURT: Are you aware of anything else? I mean, I know you said, "Well, we don't have a" -- I think you said words to the effect of, "There's not a list of every URL that crawl or data has ever visited," which if its

job is to go out into the internet everywhere, that would -- I understand that. But are you aware of any other tracking -- other than the metadata in the training sets, which we -- one hopes, but may or may not be there, then is there some other source that you're aware of?

MR. TUTTLE: I am not. The only -- I mean, the -- well, we would -- and we would have to investigate. There could be a -- like a -- an earlier version, like an earlier extraction or something like that that could contain it. If that exists, we could look for that.

THE COURT: Earlier than what? Than the training dataset?

MR. TUTTLE: Correct.

THE COURT: Uh-huh.

MR. TUTTLE: If it's -- but, again, the -- that won't -- that's not the code. It --

THE COURT: I understand.

MR. TUTTLE: -- it would be an earlier -- like a super set of the training data that existed at an earlier stage, if that still exists. That would be an option, and we could look for other ways if this comes up. I hope that it will not.

THE COURT: I hope so too.

MR. TUTTLE: But it won't be in the source code regardless.

THE COURT: Okay. All right. That is helpful. Let me get some final comments from Mr. Young, and I'll --

MR. TUTTLE: Thank you. Oh -- I'll -- no. That's it, your Honor.

THE COURT: Okay.

MR. YOUNG: Thank you, your Honor. Christopher Young on behalf of Plaintiffs. And I just have a couple of comments as I was kind of just sitting here and reflecting on the colloquy. One of the things I just want to make sure we recall -- we all collectively remember is that class membership is not the only thing we will have to show at class certification. We'll have to show that common issues predominate over individual ones. And it is our contention that the source code would be the best evidence of that, because the source code is really the best reflection of the conduct at issue, which is not just the copying, but the continual use. And I think under the case law we cited in our discovery brief, it is recognized that each use must be justified, and the source code really is the best evidence of the conduct of each infringing use, or perhaps each use that Google would say would be fair use. That's the first point.

THE COURT: But the source code is -- I mean, I always think of it as a -- not being a coder by nature, but

as a set of instructions. I mean, it's -- you know, it sets parameters, direction, and then it goes out and executes.

MR. YOUNG: Certainly.

THE COURT: I mean, it's executing an algorithm, really, right?

MR. YOUNG: Certainly. But that's exactly why it would be relevant for class cert, understanding what that common course of conduct is, especially when the documents themselves may not evidence some of the things that will be important for class cert. And by way of example, your Honor, I'll point you to -- you know, pending right now in San Francisco right now in federal court is Judge Alsup is wrestling with class certification right now in a case involving Anthropic. And, you know, he --

THE COURT: I am aware.

MR. YOUNG: Yeah. And one of the -- one of the -- one of the back and forth between the Court and the parties between the filings is whether or not there should be a pirate -- pirated works class and a non-pirated works class. And if -- what the source code will show and what we've -- at least I have not seen from the documents is any acknowledgment of whether or not Google has taken any pirated information at all. But as Mr. Tuttle has -- I think no one disagrees that Google's entire -- one of Google's main businesses is to scrape the internet. But if

they're scraping pirate websites with intent and knowledge that they're doing so to use for training the models, that could be -- that could be a dividing line for certain classes in this case.

THE COURT: And I appreciate that, and we -- I know your papers address the subclass and need for information, but if the dataset has metadata -- if the training data has the metadata, then you'll have that information.

MR. YOUNG: Potentially.

THE COURT: Yeah, ideally.

MR. YOUNG: I mean, it depends where we get the training dataset. I mean, it was unclear to me now whether we will have kind of the training data as it was fed before the model, after it gets cleaned up because -- you know, just again, it's taking a step back from the life cycle of, kind of, data, right? You have data as it's scraped, right? And I imagine there's going to be some cleanup. I mean, I do not think, for example, Google is feeding, like, URLs -- constantly feeding URLs into the model to train, because -- I just don't think from, like, a practical matter. That'll just make the thing essentially just a URL regurgitation device, and that's not how we understand these things to do, right? So if the -- I mean, I just do not know at what stage the training datasets are going to be produced to us

in a form that is -- that we're going to be inspecting. For -- so for us, which is why we're also asking for the crawlers, which at our first instance will tell us kind of the provenance of it. And we --

THE COURT: I understand.

MR. YOUNG: Yeah.

THE COURT: I understand. Okay.

MS. WEAVER: Just on the issue of the pirated and the copyrighted, if there is source code that shows filtering after the training set comes, that is very significant for the case because it tells you intent, it tells you that Google can filter copyrighted material out and chose not to do it on the front end, and it also gets at there's copyrighted material, and then there's known pirated material. And the URLs alone won't show that, and it won't get at the heart of whether they're distinguishing it at any point in the process.

It also gets at the transformative nature, you know, what are they doing with the data? What is the Court doing? We did have a proffer when we were meeting and conferring from Google to say they would not challenge this transformative element of class cert. And if we work that out and put that in a stipulation and order for the Court, maybe that could bear on that discrete issue. But there is certainly a bucket of source code that is giving

instructions about the use of categories that really go to the heart of the case.

MR. YOUNG: And, your Honor, just to put a finer point on what my colleague, Ms. Weaver, just said, you know, it is our understanding, based on our meet and confer earlier today, that Google is not going to challenge at class cert that -- the individual -- for each individual models, what happens kind of in -- at the model level, the neural network level, the conversion from training data to output would not be an individualized issue, which is why we're currently dropping, at least for class certification purposes, our ask for the code, the evidence, kind of the training and data processing. So that kind of intermediate step in between the collection, collating, and out to the last step of kind of the mitigation.

THE COURT: Okay. I appreciate that. All right. Anything further, Mr. Young?

MR. YOUNG: That's it, your Honor. Thank you.

THE COURT: Okay. All right.

All right. Thank you all very much for that. Let me take a short recess just to organize my notes and see if I have any further follow-up questions while I have everyone here. So we'll take 10 minutes and -- that might be 15, but I'm going to shoot for 10.

And Ms. (Indiscernible), if you'll come back and get me

then, I'll appreciate it.

All right. Thank you. We're in recess. You may remain seated.

(Proceedings recessed briefly.)

THE COURT: Great. Thank you for your patience. I wanted a few minutes to go back through my notes and the materials that the parties had submitted in support of their respective positions with regards to the production of source code at this stage.

I do think that the review process that's been established, there are undoubtedly some wrinkles in that, there always are in these inspection protocols, but it seems like at least it is set up to address a number of the concerns raised by Plaintiffs as to insight into the data that they need, certainly for purposes of class certification. It may not be complete, that is the datasets able to be accessed through this environment, but it's too early to tell. It looks like the objective is for it to be complete, in that the datasets -- the actual content is provided and the metadata for that content.

So with that, I don't see yet the necessity for source code, at least in the -- certainly not in the scope as the current request is pending. So I'm going to deny the request for source code without prejudice -- I want -- provided that Plaintiffs' access to the datasets in the

environment that's been described here today by Google is available next week. And I know it's going to take some time to get oriented and start to look at that and evaluate what is there, and I'm -- obviously, it's in Plaintiffs' best interest to jump on that and move through that. And if there are data points that go to issues of class certification that are missing or cannot be discerned, you need to meet and confer on those. But this has to be prompt. We're all on the clock here. And response to that has to be prompt. And if there is, again, something that can be identified or a piece of necessary data for class cert that can be identified that's not available through this environment, inspection of datasets, and -- then Plaintiffs can come back with a request for the source code, and it should -- it would be more limited and focused on what is missing. But I'm hopeful that through the meet and confer process, the parties could come to some agreement.

It may well be that at some point in this process that source code has to be produced and made available. It's not unusual in these cases. But let's get a look at what of the needs are met through the dataset inspection as it is set up. So that'll get started next week.

If Plaintiffs have identified a dataset that Google cannot confirm was used in training -- and I understand your point, Mr. Tuttle, that to the extent datasets are picked

off of data cards, they were used in training. But if they come from some other source, if Google cannot confirm that they were used in training, then that needs to be communicated promptly here at the start of the review process, and the Plaintiffs can replace that dataset with some other selection, because obviously we need -- you need the information that was actually used in training.

All right. So I am going to issue a -- I'll issue a summary order from today that captures these points with regards to both source code and the custodian issue, so everyone's record is clear and the docket is clear. And -- but my rulings are as I have articulated them from the bench here today.

All right. Any questions? Anything that I've just said that you're thinking, "Oh, no, that can't possibly happen." And hearing none, excellent. Thank you.

I want to, again, thank Counsel for your preparation, both in the papers and the supporting materials for today's positions -- presentations, as well as the presentations themselves, and for restating your names every time you were at the podium for the benefit of my clerks.

All right. Thank you all very much. Have a good day.

MR. YOUNG: Thank you, your Honor.

MS. WEAVER: Thank you, your Honor.

MR. TUTTLE: Thank you, your Honor.

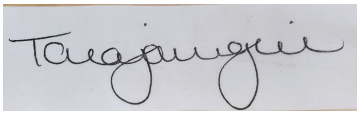
THE CLERK: We are adjourned.

(Proceedings adjourned at 11:49 a.m.)

CERTIFICATE OF TRANSCRIBER

I certify that the foregoing is a true and correct transcript, to the best of my ability, of the above pages of the official electronic sound recording provided to me by the U.S. District Court, Northern District of California, of the proceedings taken on the date and time previously stated in the above matter.

I further certify that I am neither counsel for, related to, nor employed by any of the parties to the action in which this hearing was taken; and, further, that I am not financially nor otherwise interested in the outcome of the action.



Echo Reporting, Inc., Transcriber

Monday, June 23, 2025